第 25 卷第 1 期
2008 年 1 月
统计研究
Statistical Research
Vol. 25, No. 1
Jan. 2008

数据挖掘中一种新的聚类方法*

——基于对应分析与因子旋转

殷瑞飞　朱建平

内容提要:本文基于 Q 型因子分析的基本思想,结合对应分析方法,建立了一种适用于大型数据库聚类的方法。该方法既解决了 Q 型因子分析算法效率方面的问题,也解决了传统对应分析法中缺乏客观分类标准、信息损失严重等多种缺陷,在实证分析中也取得了良好的效果。

关键词:数据挖掘;聚类分析;对应分析;因子旋转
中图分类号:C812　　**文献标识码**:A　　**文章编号**:1002 - 4565(2008)01 - 0093 - 05

A New Clustering Method in Data Mining

——Based on Correspondence Analysis and Factor Rotation

Yin Ruifei & Zhu Jianping

Abstract: Based on the idea of Q-mode factor analysis and correspondence analysis, this paper proposes a new clustering approach to fit for large-scaled database. The approach is effective in calculation which is an obstacle in Q-mode factor analysis. Additionally, this approach overcomes the subjectivity of traditional correspondence analysis and avoids the lost of information. The validity of the proposed approach is verified with a case.

Key words: Data mining; Cluster analysis; Correspondence analysis; Factor rotation

一、引言

数据挖掘(Data Mining)是近几年随着数据库和人工智能发展起来的一门新兴技术,它从大量原始数据中发掘出隐含的、有用的信息和知识,帮助决策者寻找数据间潜在的关联,发现被忽略的因素。

本文所述聚类方法的思想来源于 Q 型因子分析[8-10]。因此,该方法并不能直接应用到数据挖掘领域。Guttman (1941) 以内部一致性 (internal consistency) 作为计算准则[11],Benzécri (1969) 基于对数据矩阵的重新标度,分别从不同角度证明了对应分析中 R 型因子解和 Q 型因子解的对偶性,从而为解决 Q 型因子分析运算速度上的瓶颈提供了思路。本文正是基于 Benzécri 对应分析的基本思路,在卡方距离框架下构造了一个与经典 Q 型因子分析类似的因子载荷阵,并给出了因子得分的求解方法。另外,本文将因子旋转引入对应分析,通过对 Q 型因子载荷阵进行方差最大旋转,使得聚类结果更加清晰。通过算法分析,该方法的时间复杂度是样本容量的线性阶,这充分体现了其在算法效率上的优越性。同时,该方法可以解决传统对应分析法中诸如缺乏客观分类标准、信息损失严重等多种缺陷。利用该方法对移动通讯月度消费大型数据库进行聚类分析也取得了较好的效果。

二、用因子载荷对样本聚类的基本思想

对于由 n 个样品和 p 个变量构成的 $n \times p$ 初始数据矩阵,Q 型因子分析就是把 n 个样品分别表示为 k 个公共因子和一个特殊因子的线性加权和。即:

$$O_i = a_{i1} F_1 + a_{i2} F_2 + \cdots + a_{ik} F_k + \varepsilon_i \quad i = 1, 2, \cdots, n \tag{1}$$

用矩阵表示为:

$$\underset{(n\times 1)}{O} = \underset{(n\times k)}{A}\ \underset{(k\times 1)}{F} + \underset{(n\times 1)}{\varepsilon} \tag{2}$$

*国家教育部新世纪优秀人才支持计划(NCET-04-0608)资助;国家教育部社科研究规划项目(06JA910003)资助。

其中,O_i 表示第 i 个随机样品,F_1,F_2,…,F_k 为公共因子,ε_i 为特殊因子。O 是由 n 个样品构成的随机向量。矩阵 A 称为因子载荷阵,系数 a_{ij} 称为因子载荷,它表示了第 i 个样品对第 j 个公共因子的相对依赖程度。可以证明因子载荷阵满足如下两个性质:

性质 1:因子载荷阵 A 就是随机向量 O 和 F 的协差阵。即

$$Cov(O,F) = A \tag{3}$$

如果数据是经过行标准化的,则矩阵 A 就是随机向量 O 和 F 的相似系数矩阵,a_{ij} 就是 O_i 和 F_j 的相似系数。

性质 2:样品 O_i 的方差可以分解为公因子解释的方差和特殊因子方差两部分。即

$$Var(O_i) = a_{i1}^2 + a_{i2}^2 + \cdots + a_{ik}^2 + \sigma_i^2$$

当提取的因子数目足够多时,特殊因子方差可以忽略,此时有:

$$Var(O_i) \approx a_{i1}^2 + a_{i2}^2 + \cdots + a_{ik}^2 \tag{4}$$

从空间几何的角度来看,在 k 维因子空间中,样品 O_i 可以用向量 $O_i = (a_{i1}, a_{i2}, \cdots, a_{ik})$ 表示,并且对应于因子空间中的一个点。对于标准化过的数据,有 $Var(O_i) = 1$。这样,向量 O_i 的长度 $\| Q_i \| = \sum_{j=1}^{k} a_{ij}^2 \approx 1$,即因子空间中的样品点都大致落在单位超球面上,如图 1 所示。

利用因子载荷阵对样本聚类的基本思想就是通过判断不同样品在各个因子上载荷的相对大小来构造分类标准:如果某几个样品在同一个因子上都有相对较大的正载荷,则说明这几个样品与该因子同时具有较强的正相似性,于是这几个样品就可以聚为一类;相反,如果某几个样品在同一个因子上都有较大的负载荷,则说明这几个样品与该因子同时具有较强的负相似性,于是这几个样品可以聚为另一类。这样,如果提取 k 个因子,样品将被分为 $2k$ 类。图 1 中,在 F_1 上有较大正载荷的 Q_1 和 Q_2 归为一类,在 F_2 上有较大正载荷的 Q_3 和 Q_4 归为一类,在 F_2 上有较大负载荷的 Q_5、Q_6 和 Q_7 归为一类。

如果因子载荷阵中各个样品对不同因子的载荷的绝对值差别不是很明显,则可以通过对载荷阵进行方差最大旋转,使每个样品仅在一个公共因子上有绝对值较大的载荷,而在其他因子上载荷的绝对值较小,从而使样品分类清晰化。

然而,通过算法分析可以知道,若对一个由 n 个样品和 p 个变量构成的数据矩阵进行 Q 型因子分析,必须要求一个 $n \times n$ 协差阵的特征根及相应的特征向量,其算法的时间复杂度为 $O(n^3)$,也就是说其运行时间大致与 n^3 成正比。那么,当样本容量 n 很大的时候,Q 型因子分析将耗费大量的计算时间,因此,该方法并不能直接应用于数据挖掘领域。

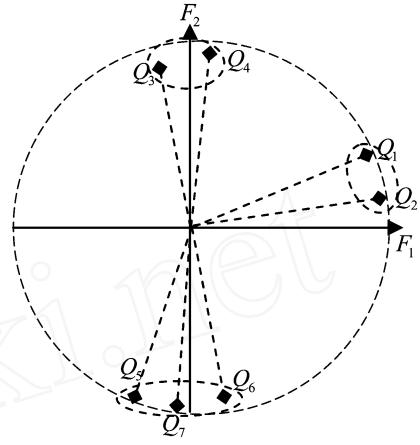

图 1 因子载荷图

三、对应分析聚类法

为了解决上述算法效率方面的缺陷,首先借助了 Benzécri 对应分析的思路,利用对应分析中 R 型因子解和 Q 型因子解之间的对偶性,由 R 型因子解间接地得到 Q 型因子载荷阵。然后将对应分析中得到的因子载荷进行适当变换,使其能够真正地代表样品与因子的相似程度。这时,就可以利用因子载荷作为分类标准对样本进行聚类,把这种方法称为"对应分析聚类法"。

(一)Benzécri 对应分析

Benzécri 对应分析的最大特点是对原始数据的重新标度,即将原始数据从欧式距离空间转换到卡方距离空间。设原始数据矩阵 $X = (x_{ij})_{n \times p}$。首先,对原始数据矩阵进行变换,得到过渡矩阵 $Z = (z_{ij})_{n \times p}$。其中

$$z_{ij} = \frac{x_{ij} - x_{i.}\, x_{.j} / x_{..}}{\sqrt{x_{i.}\, x_{.j}}}$$

则可以证明,p 阶方阵 $\sum_c = Z'Z$ 就是卡方距离意义下 p 个变量点的协差阵,而 n 阶方阵 $\sum_r = ZZ'$ 就是卡方距离意义下 n 个样品点的协差阵,且 $\sum_r$ 和

$\sum_c$ 具有完全相同的非零特征根 $\lambda_1, \lambda_2, \cdots, \lambda_k$ ($k \leq \min\{n, p\} - 1$)。同时可以证明,若 u_j 为特征根 λ_j 关于 $\sum_c$ 的特征向量,则特征根 λ_j 关于 $\sum_r$ 的特征向量 $v_j = Zu_j$。

这样,就建立了对应分析中 R 型因子分析和 Q 型因子分析的关系。也就是说,不再需要 n 对阶方阵 $\sum_r$ 求特征根和向量,而可以从 p 阶方阵 $\sum_c$ 出发直接得到 R 型因子载荷阵

$$B = (u_1 \sqrt{\lambda_1}, u_2 \sqrt{\lambda_2}, \cdots, u_k \sqrt{\lambda_k}) \tag{5}$$

以及 Q 型因子载荷阵

$$A = (v_1 \sqrt{\lambda_1}, v_2 \sqrt{\lambda_2}, \cdots, v_k \sqrt{\lambda_k}) \tag{6}$$

从而解决了 Q 型因子分析运算速度上的瓶颈。①

(二)因子载荷阵的标准化

如前所述,矩阵 $\sum_c = Z'Z$ 和矩阵 $\sum_r = ZZ'$ 分别代表了卡方距离意义下变量点的协差阵和样品点的协差阵。因此有与式(3)类似的结论:Q 型因子载荷 a_{ij} 代表了样品 Q_i 与因子 F_j 在卡方距离意义下的协方差,即

$$a_{ij} = \mathrm{Cov}_{\chi^2}(O_i, F_j) \text{ ②} \tag{7}$$

为了得到一个能够代表样品与因子之间相似系数的因子载荷阵,需要对式(6)中的因子载荷阵 A 进行一定的加工。

矩阵 $\sum_r$ 的第 i 个对角线元素 $\sum_{j=1}^{p} z_{ij}^2$ 就是样品 O_i 在卡方距离意义下的方差,记为 $\mathrm{Var}_{\chi^2}(O_i)$。因此,对载荷阵 A 的第 i 行除以样品 O_i 的标准差 $\sqrt{\sum_{j=1}^{p} z_{ij}^2}$ ($i = 1, 2, \cdots, n$),得到一个标准化的因子载荷阵 A^*,即

$$A^* = (a_{ij}^*)_{n\times k} = \left(\mathrm{diag}\sum_r\right)^{-1/2} A \tag{8}$$

$$a_{ij}^* = \frac{a_{ij}}{\sqrt{\sum_{j=1}^{p} z_{ij}^2}} \tag{9}$$

则有

$$r_{\chi^2 O_i, F_j} = \frac{\mathrm{Cov}_{\chi^2}(O_i, F_j)}{\sqrt{\mathrm{Var}_{\chi^2}(O_i)\mathrm{Var}_{\chi^2}(F_j)}} = \frac{a_{ij}}{\sqrt{\sum_{j=1}^{p} z_{ij}^2}} = a_{ij}^* \tag{10}$$

即 a_{ij}^* 代表了样品 O_i 与因子 F_j 在卡方距离意义下的相似系数。

另外,与式(4)类似的有:$\sum_{j=1}^{k} a_{ij}^2 \approx \mathrm{Var}_{\chi^2}(O_i)$。从而新的因子空间中向量 O_i^* 的长度:

$$\| Q_i^* \| = \sum_{j=1}^{k} a_{i1}^{*2} = \sum_{j=1}^{k} \left(\frac{a_{ij}}{\sqrt{\sum_{j=1}^{p} z_{ij}^2}} \right)^2$$

$$= \frac{\sum_{j=1}^{k} a_{ij}^2}{\sum_{j=1}^{p} z_{ij}^2} = \frac{\sum_{j=1}^{k} a_{ij}^2}{\mathrm{Var}(O_i)} \approx 1 \tag{11}$$

即新的载荷阵中代表样品点的向量在因子空间中均落在单位超球面附近。

这样,就在卡方距离的框架下构造了一个与经典 Q 型因子分析类似的因子载荷阵。接下来,就可以利用因子载荷的相对大小作为分类标准对样本进行聚类。

为了使样品分类更加清晰,可以对因子载荷阵 A^* 进行方差最大旋转③。即寻找一个旋转矩阵 T,使得旋转后载荷阵 $A_r^* = A^* T$ 每一列元素的绝对值向 0、1 两极分化,以使得各个样品更密集地聚集在不同的因子轴附近。

(三)因子得分与类的解释

当聚类完成之后,人们经常利用类重心作为类的代表,用类重心的坐标来解释各类的特征。在 Q 型因子分析中,所有的样品被分别聚在 k 个因子轴的周围,因此可以用公因子来代表各类,用 Q 型因子得分来解释各类的特征。同样地,在对应分析聚类法中,也可以计算 Q 型因子得分,从而对各类的特征做出描述。

估计因子得分有很多方法,常用的有回归法和加权最小二乘法[14]。这里采用加权最小二乘法来

① 在实际应用中,人们常常直接利用对应分析的结果,将样品点和变量点共同画在一张二维散点图上,通过观察各散点的相对位置来判断样品之间的亲疏程度,从而达到聚类的目的。这种方法至少存在两大缺陷:一是通过主观观察得出结论,缺乏客观的分类标准;二是将 k 维因子载荷投影到 2 维空间中,经常导致严重的信息损失。

② 为了避免与前文符号混淆,本文将卡方距离意义下的方差、协方差、均值等均以下标 χ^2 标注。

③ 我们的最终目的是使得分类结果更加清晰化,而并不在意旋转后的因子是否相互独立,因此,理论上来讲,使用斜交旋转将得到更好的结果。但是,经过多次试验,我们发现使用斜交旋转与正交旋转对聚类结果并无显著影响。因此,本文采用正交因子旋转法。

估计因子得分①:

$$f_{(j)} = (A'A)^{-1}A'(X_{\chi^2(j)} - \bar{X}_{\chi^2}) \quad (j = 1,2,\cdots,p) \quad (12)$$

其中,A 为 Q 型因子载荷阵,$X_{\chi^2(j)}$ 和 $\bar{X}_{\chi^2}$ 分别为卡方距离框架之下的样品点向量和样品点均值向量。即

$$X_{\chi^2(j)} = \left(\frac{p_{1j}}{p_{\cdot j}\sqrt{p_{1\cdot}}}, \frac{p_{2j}}{p_{\cdot j}\sqrt{p_{2\cdot}}}, \cdots, \frac{p_{nj}}{p_{\cdot j}\sqrt{p_{n\cdot}}}\right)',$$

$$\bar{X}_{\chi^2} = (\sqrt{p_{1\cdot}}, \sqrt{p_{2\cdot}}, \cdots, \sqrt{p_{n\cdot}})'$$

其中,

$$p_{ij} = x_{ij} \Big/ \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}, \quad p_{i\cdot} = \sum_{j=1}^{p} p_{ij}, \quad p_{\cdot j} = \sum_{i=1}^{n} p_{ij}$$

在对应分析聚类法中,对类别特征做出解释的另外一种更为直接的办法是利用对应分析将样本和变量置于同一空间的特性,直接观察旋转后的 R 型因子载荷②。若某变量在一个因子上有较大载荷,则说明该因子所代表的类在该变量上有较大取值。

(四)算法分析

评价算法效率的常用指标是算法的时间复杂度,即随着"问题规模"的增长而导致的算法执行时间的增长率。所谓"问题规模",是指算法中输入量的数目。表 1 给出了对应分析聚类算法的基本步骤以及每一步骤的时间复杂度。其中,n 为样本容量,p 为变量个数,k 为提取的因子数,i 为提取因子时的迭代次数,r 为因子旋转的迭代次数③。

表 1 对应分析聚类法的时间复杂度

	算法步骤	时间复杂度
1	计算过渡矩阵 Z	$O(np)$
2	计算 p 阶方阵 $\sum_c = Z'Z$	$O(np^2)$
3	求解 R 型因子载荷	$O(ip^3)$
4	利用 R 型解得到 Q 型解	$O(npk)$
5	对 Q 型因子载荷阵进行旋转	$O(nrk^2)$
6	计算 Q 型因子得分	$O(npk)$

当样本容量相对于其他输入量很大时,其他输入量均可忽略,从而,对应分析聚类法总的时间复杂度为:

$$O(np) + O(np^2) + O(ip^3) + O(npk) + O(nrk^2) + O(npk) = O(n)$$

这说明对应分析聚类法的时间复杂度是样本容量的线性阶,这充分体现了该方法在算法效率上的优越性。

四、实证分析

笔者使用对应分析聚类法对一个包含约 10 万条记录的移动通讯月度消费数据进行分析。参与分析的变量共有 5 个,分别为市话费、长话费、漫游费、信息费和新业务费④。这里提取 3 个因子,从而将 10 万名手机用户分为 6 类。

表 2 为旋转后 Q 型因子载荷阵。从该表中可以清晰地看到不同手机用户的聚类结果:在因子 1 上具有较大正载荷的用户被分到第 1 类,具有较大负载荷的被分到第 2 类;在因子 2 上具有较大正载荷的被分到第 3 类,具有较大负载荷的被分到第 4 类;在因子 3 上具有较大正载荷的被分到第 5 类,具有较大负载荷的被分到第 6 类。

表 2 旋转后 Q型因子载荷阵(截选)

类别	用户号码	因子 1	因子 2	因子 3	最大载荷序号	最大载荷符号
1	139 * * * *0288	0.901	0.004	0.434	1	+
	139 * * * *9359	0.903	- 0.012	0.214	1	+
	138 * * * *9857	0.669	0.174	0.458	1	+
	…	…	…	…	…	…
2	135 * * * *5712	- 0.916	- 0.287	0.168	1	-
	138 * * * *9997	- 0.842	- 0.523	- 0.119	1	-
	139 * * * *6027	- 0.811	- 0.366	0.175	1	-
	…	…	…	…	…	…
3	138 * * * *6198	- 0.351	0.926	- 0.140	2	+
	138 * * * *7820	0.054	0.995	- 0.086	2	+
	138 * * * *3228	- 0.159	0.819	- 0.499	2	+
	…	…	…	…	…	…
4	138 * * * *5656	- 0.308	- 0.912	- 0.217	2	-
	138 * * * *4766	0.362	- 0.839	- 0.237	2	-
	139 * * * *1533	0.097	- 0.954	- 0.244	2	-
	…	…	…	…	…	…
5	139 * * * *1990	- 0.080	0.554	0.826	3	+
	139 * * * *8959	- 0.141	0.368	0.918	3	+
	139 * * * *3326	- 0.569	0.085	0.817	3	+
	…	…	…	…	…	…
6	138 * * * *8195	- 0.154	0.519	- 0.793	3	-
	135 * * * *9515	0.485	0.203	- 0.846	3	-
	138 * * * *7388	- 0.165	- 0.333	- 0.927	3	-
	…	…	…	…	…	…

表 3 为旋转后 R 型因子载荷阵。表 4 为各类用

① 在回归法中,需要对协差阵求逆,而由对应分析的理论可知,样本协差阵 $\sum_r$ 是一个降秩矩阵,因此,这里无法使用回归法估计因子得分。

② 值得注意的是,这里并非直接针对 R 型因子载荷进行方差最大旋转。此处的"旋转后 R 型因子载荷"是标准化后的 R 型因子载荷在上述 Q 型因子旋转之后得到的因子空间中的坐标。即首先对初始 R 型因子载荷阵 B 每一行除以卡方距离意义下变量点标准差,得到一个标准化的 R 型因子载荷阵 B^*,然后对其直接右乘上述 Q 型因子旋转矩阵 T。

③ 经验表明,提取因子时的迭代次数和因子旋转的迭代次数均不大,一般在不超过 30 次迭代后达到收敛。

④ 新业务是指移动秘书、呼叫转移、手机上网、彩信、语音杂志等特殊业务。

户在各种话费上的均值。结构表 3 —表 5 的信息可以清晰地看出 6 类不同用户群体的消费特征。因子 1 正方向所代表的第 1 类用户群体的消费特征为高漫游费、低市话费,该类用户月平均市话费仅不到 30 元,漫游费却平均高达 71.5 元,可将此类用户命名为"商务型用户"。因子 1 负方向所代表的第 2 类用户恰好与第 1 类相反,其特征是高市话费、低漫游费,其月平均市话费为 47.6 元,平均漫游费仅为 1.1 元,可将此类命名为"本地型用户"。

第 3 类用户群体特征为新业务费比例较高,月平均为 22.4 元,几乎接近市话费的水平,可命名为"新兴型用户"。与之相反的是第 4 类用户,他们几乎不使用新业务,新业务费月平均不到 1 元,称其为"传统型用户"。

第 5 类是信息费比例较高的用户群体,月平均信息费为 15.8。与之相反的是第 6 类用户,其每月平均信息费仅 1.1 元。

有了分类之后,就可以根据不同用户群体的消费倾向制定有针对性的营销策略。

表 3　　旋转后 R型因子载荷阵

变量	因子 1	因子 2	因子 3
市话	- 0.972	- 0.002	0.058
长话	- 0.162	- 0.058	- 0.098
漫游	0.770	- 0.462	- 0.436
信息	0.372	0.030	0.928
新业务	0.310	0.930	- 0.196

表 4　　不同用户群体在各种话费上的均值

	第 1 类	第 2 类	第 3 类	第 4 类	第 5 类	第 6 类
市话	29.4	47.6	24.5	65.9	18.2	47.5
长话	5.3	2.9	2.3	8.6	1.5	6.5
漫游	71.5	1.1	5.4	18.5	3.9	10.9
信息	6.0	2.9	6.9	2.6	15.8	1.1
新业务	5.3	1.6	22.4	0.9	1.1	12.0
样本容量	15893	60176	13780	10484	21141	4878

五、结语

本文基于 Q 型因子分析的基本思想,结合对应分析方法,在卡方距离框架下建立了一种新的大型数据库聚类方法。该方法既解决了 Q 型因子分析算法效率方面的问题,也解决了传统对应分析法中缺乏客观分类标准、信息损失严重等多种缺陷。

当然,本文的方法远非完美。例如,若原始数据有 p 个变量,则对应分析聚类法将最多可以将其分为 $2(p-1)$ 类。因此,如何扩大可能的分类数,将是进一步研究的一个重点。另外,即使经过因子旋转,依然会有大量样品在不同因子上载荷的绝对值差异不够显著,从而使得分类界限模糊。针对这一问题,目前有两种思想:一是把变量共同度小于一定标准的样品作为异常点提取出来单独进行分析;二是将本文的方法与模糊聚类方法相结合,构造一种基于模糊逻辑的聚类方法。

参考文献

[1] T. Zhang, R. Ramakrishnan, M. Livny. BIRCH: An efficient data clustering method for very large databases[C]. Proc. ACM-SIGMOD Int. Conf. Magement of data (SIGMOD'96), 1996: 103—114.

[2] S. Guha, R. Rastogi, K. Shim. CURE: an efficient clustering algorithm for large database[J]. Information Systems, 2001, 26(1): 35—58.

[3] J. Macqueen. Some methods for classification and analysis of multivariate observations[C]. Proc. 5th Berkeley Symp. Math. Statist, 1967, 1:281—297.

[4] L. Kaufman, P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis[M]. New York: John Wiley & Sons, 1990.

[5] M. Ester, H. P. Kriegel, J. Sander, X. Xu.. A density-based algorithm for discovering clusters in large spatial databases[C]. Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'96), 1996: 226—231.

[6] W. Wang, J. Yang, R. Muntz. STING: A statistical information grid approach to spatial data mining[C]. Proc. Int. Conf. Very Large Data Bases (VLDB'97), 1997:186—195.

[7] J. W. Shavlik, T. G. Dietterich, Readings in Machine Learning[R], San Mateo, CA: Morgan Kaufmann, 1990.

[8] Cattell Raymond B. Factor Analysis[M]. New York: Harper and Brothers, 1952: 462.

[9] Cattell Raymond B. The Three Basic Factor Analytic Research Designs—Their Interrelations and Derivatives[J]. Psychological Bulletin, 1952, 49: 499—520.

[10] Stephenson, William. Some Observations on Q Technique[J]. Psychological Bulletin, 1952, 49: 483—498.

[11] Guttman, L. The quantification of a class of attributes: A theory and Method of scale construction[C]. The Committee on Social Adjustment (ed.), The Prediction of Personal Adjustment. New York: Social Science Research Council, 1941.

作者简介

殷瑞飞,男,山西长治人,1980 年生,厦门大学计划统计系博士研究生,研究方向:多元统计理论与方法、数据挖掘。

朱建平,男,河南浚县人,1962 年生,2003 年毕业于南开大学数学科学学院统计学系,获得理学博士学位,现为厦门大学经济学院和王亚南经济研究院教授、博士生导师、计划统计系副主任,主要研究方向:数理统计、数据挖掘。

(责任编辑:李峻浩)